

Lustre – the future

Peter Braam

CFS Cluster
File
Systems, Inc.

lustre™

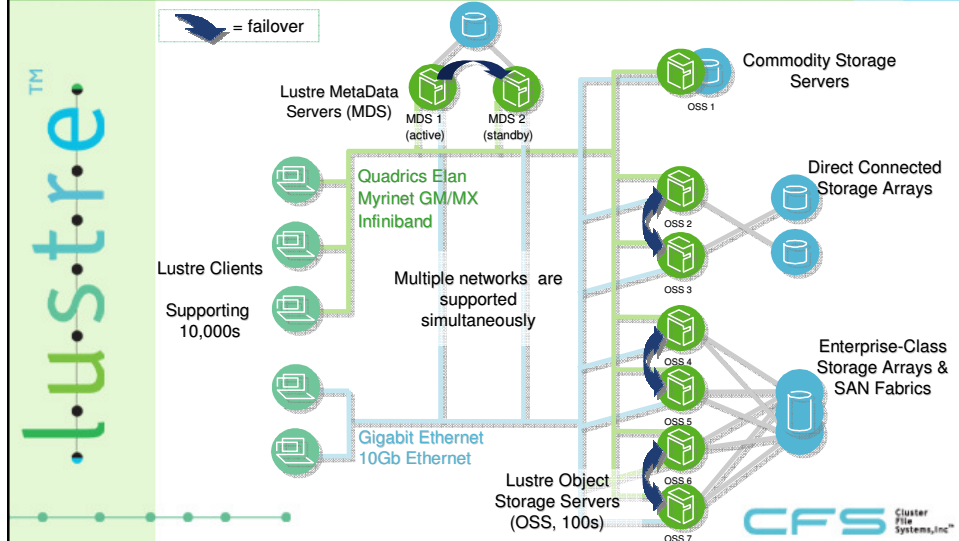
Lustre Today

lustre™

Scalable to 10,000's of CPUs	Scalable architecture provides users more than enough headroom for future growth - simple incremental HW addition
Supports Multiple Network Protocols	Support for native high-performance network IO guarantees <i>fast</i> client & server side I/O Routing enables multi-cluster, multi-network namespace
SW-only Platform	HW agnosticism allows users flexibility of choice - eliminates vendor lock-ins
Proven Capability	Experience in multiple industries; 100s of deployed systems demonstrates production-quality that can exceed 99% uptime
Failover Capable	Lustre's failover protocols eliminate single-points-of-failure and increase system reliability
POSIX Compliant	Users benefit from strict semantics and data coherency
Broad Industry Support	With the support of many leading OEMs and a robust user base, Lustre is an emerging file system standard with a very promising future

CFS Cluster
File
Systems, Inc.™

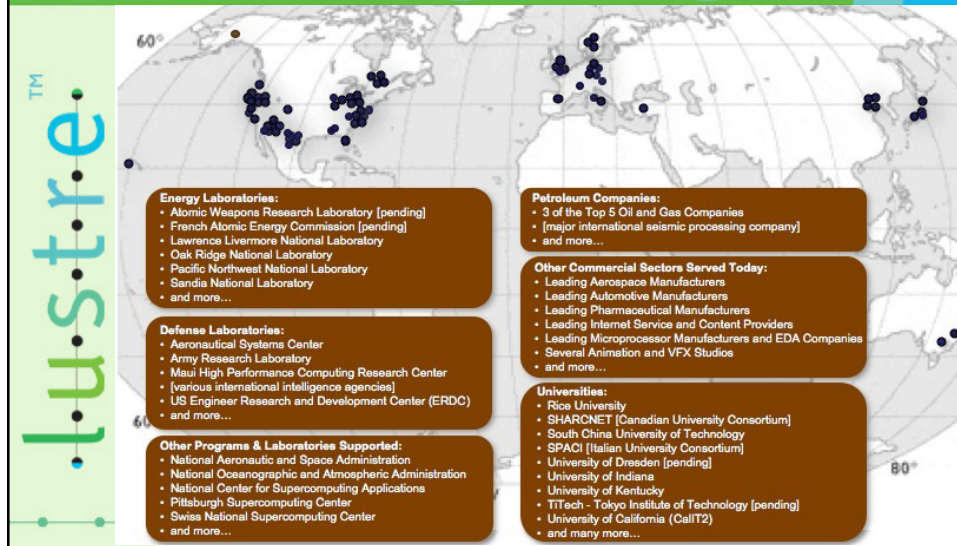
Lustre Architecture Today



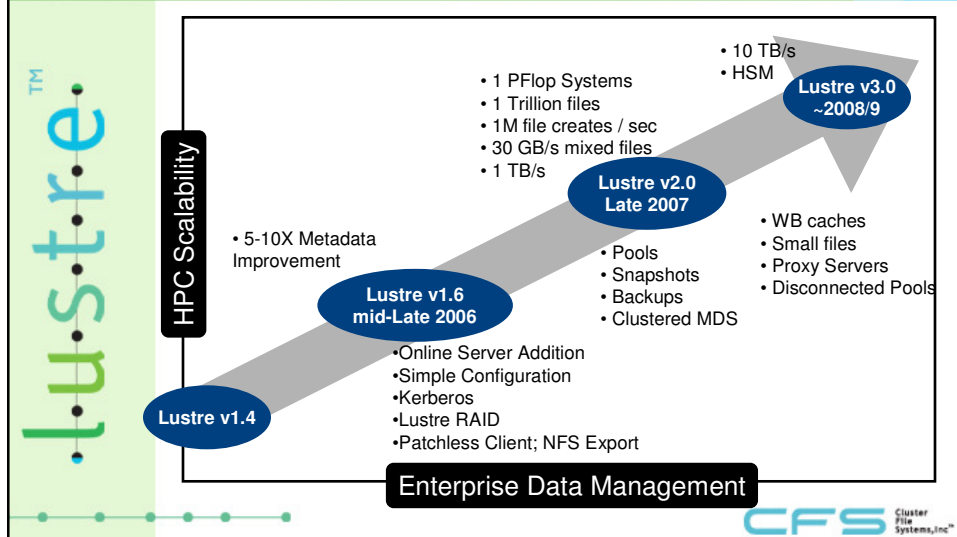
Lustre Today: version 1.4.6

Scalability	Clients: 10,000 / 130,000 CPUs Object Storage Servers: 256 OSSs MetaData Servers: 1 (+ failover) Files: 500M File System:
Performance	Single Server: 2 GB/s + Single Client: 2 GB/s + Single Namespace: 100 GB/s File Operations: 9000 ops/second
Redundancy	No Single Point of failure with shared storage
Security	POSIX ACLs
Management	Quota
Kernel/OS Support	RHEL 3, RHEL 4, SLES 9, Patches required Limited NFS Support; Full CIFS Support
POSIX Compliant	Yes

User Profile: Primarily High-End HPC



Intergalactic Strategy



Near & mid term

lustreTM

CFS Cluster File Systems, Inc.TM

12 month strategy

lustreTM

- Beginning to see much higher quality
- Adding desirable features and improvements
 - Some re-writes (recovery)
- Focus largely on doing existing installations better
- Much improved Manual
- Extended training
- More community activity

CFS Cluster File Systems, Inc.TM

In testing now



- Improved NFS performance (1.4.7)
 - Support for NFSv4 (pNFS when available)
 - Broader support for data center HW; Site-wide storage enablement
- Networking (1.4.7, 1.4.8)
 - OpenFabrics (fmr. OpenIB Gen 2) Support,
 - CISCO I/B,
 - Myricom MX
- Much simpler configuration (1.6.0)
- Clients without patches (1.4.8)
- Ext3 large partition support (1.4.8)
- Linux software RAID5 improvements (1.6.0)



More 2006



- Metadata Improvements (several ... from 1.4.7)
 - Client-side Metadata & OSS read-ahead
 - More parallel operations
- Recovery improvements
 - Move from timeouts to health detection
 - More recovery situations covered
- Kerberos5 Authentication, OSS capabilities
- Remote UID Handling
 - WAN: Lustre Awarded SC|05 for long-distance efficiency; 5GB/s over 200 miles on HP system



Late 2006 / early 2007

lustre™

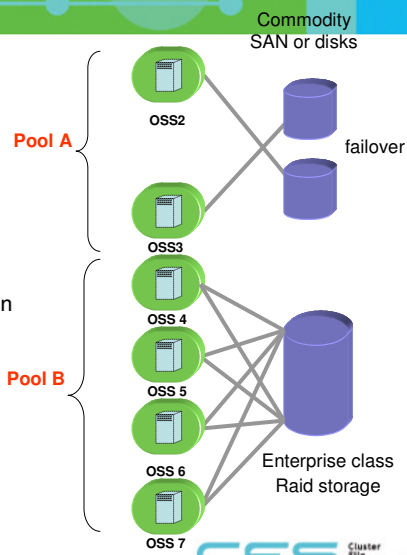
- Manageability
 - Backups – restore files with exact or similar striping
 - Snapshots – use LVM with distributed control
 - Storage Pools
- Trouble-shooting
 - Better error messages
 - Error analysis tools
- High Speed CIFS Services – pCIFS Windows client
- OS X Tiger client – no patches to OS X
- Lustre Network RAID 1 and RAID 5 across OSS Stripes

CFS Cluster File Systems, Inc.™

Pools – naming a group of OSS's

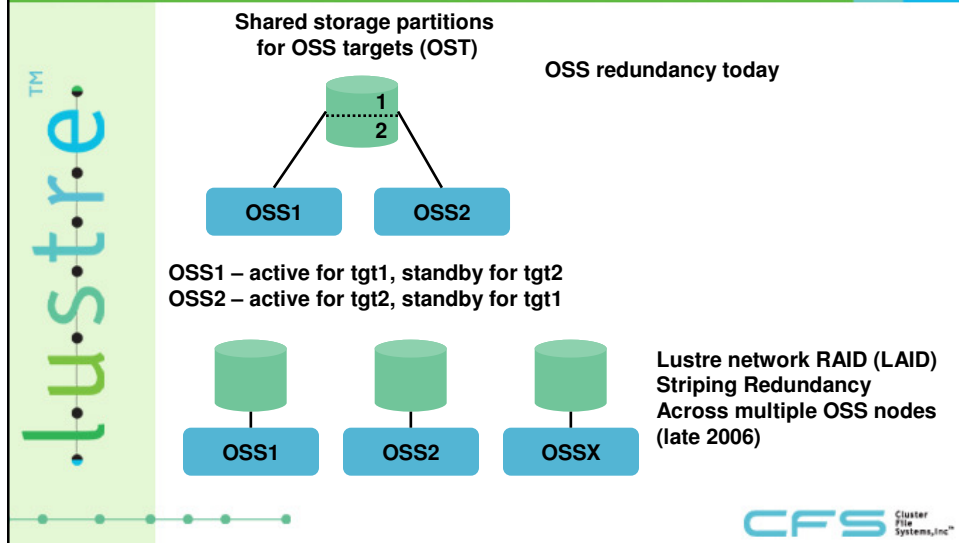
lustre™

- Heterogeneous servers
 - Old and New OSS's
- Homogeneous QOS
 - Stripe inside pools
- Pools – name an OST group
 - Associate qualities with pool
- Policies
 - Subdirectory must place stripes in pool (setstripe)
 - A Lustre network must use one pool (lctl)
 - Destination pool depends on file extension
 - Keep LAID stripes in pool

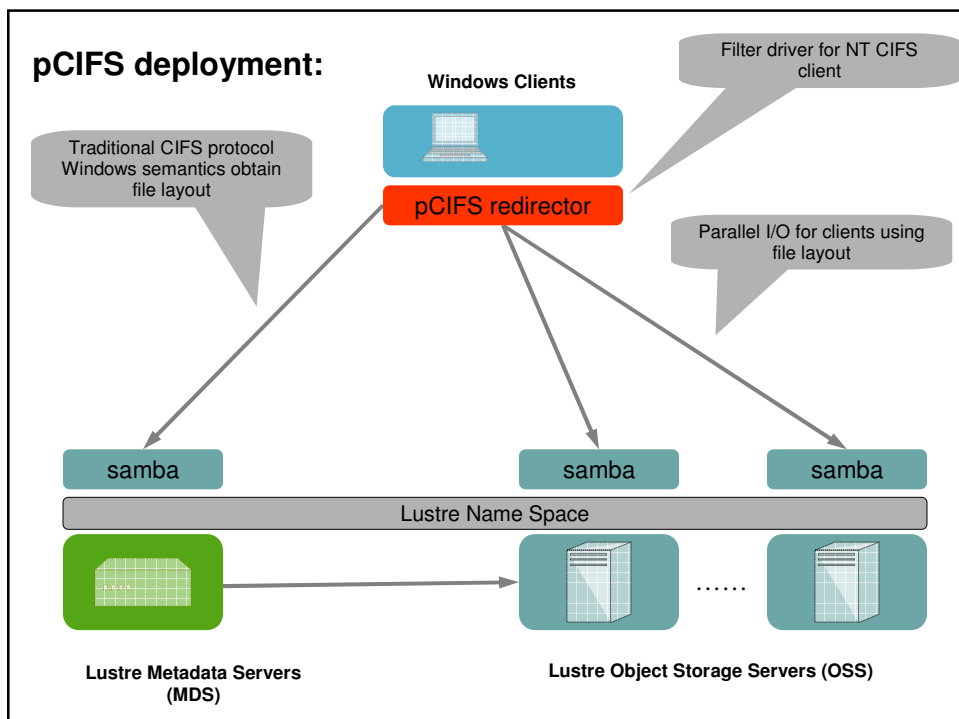


CFS Cluster File Systems, Inc.™

Lustre Network RAID



pCIFS deployment:



Longer term

lustre™

CFS Cluster File Systems, Inc™

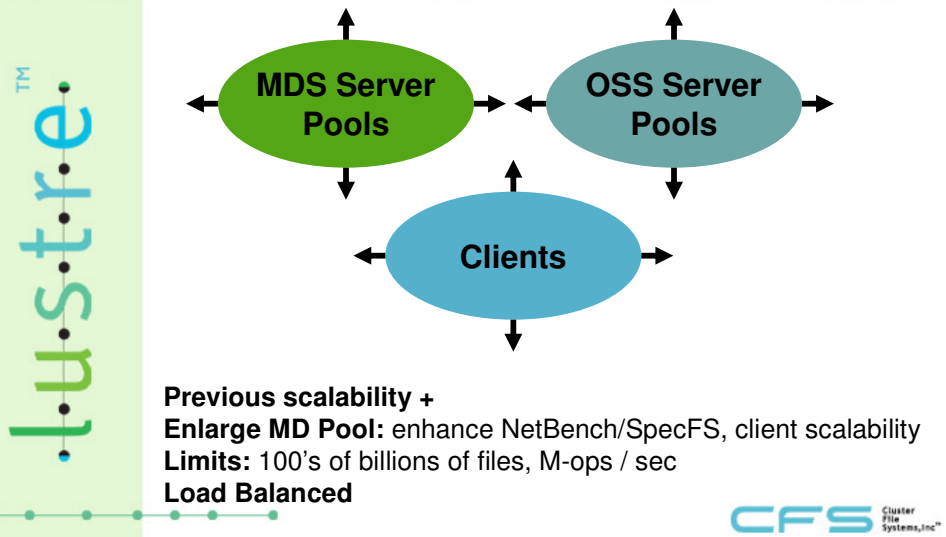
Future Enhancements

lustre™

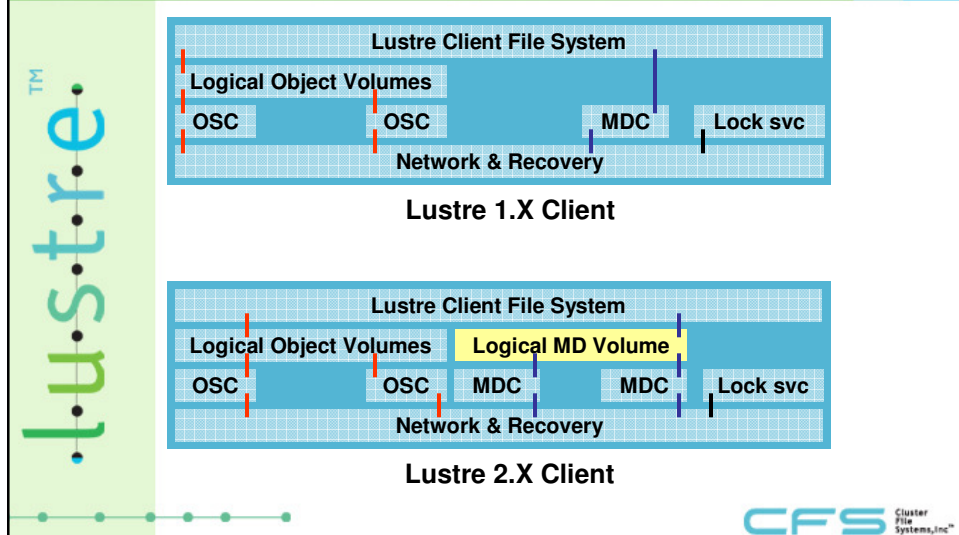
- Clustered metadata – towards 1M pure MD ops / sec
- FSCK free file systems – immense capacity
- Client memory MD WB cache – smoking performance
- Hot data migration – HSM & management
- Global namespace
- Proxy clusters – wide area grid applications
 - Disconnected operation

CFS Cluster File Systems, Inc™

Clustered Metadata: v2.0

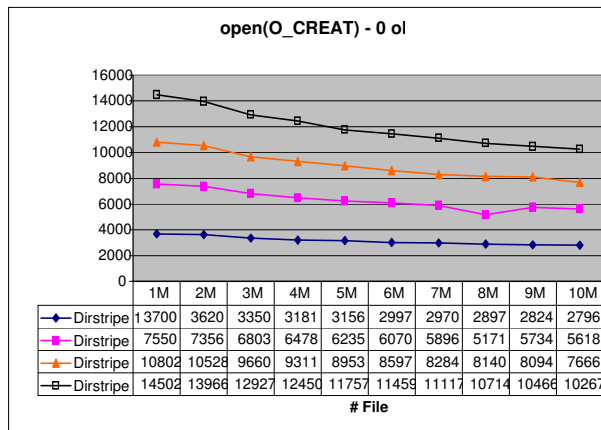


Clustered Metadata Modular Architecture



Preliminary Results - creations

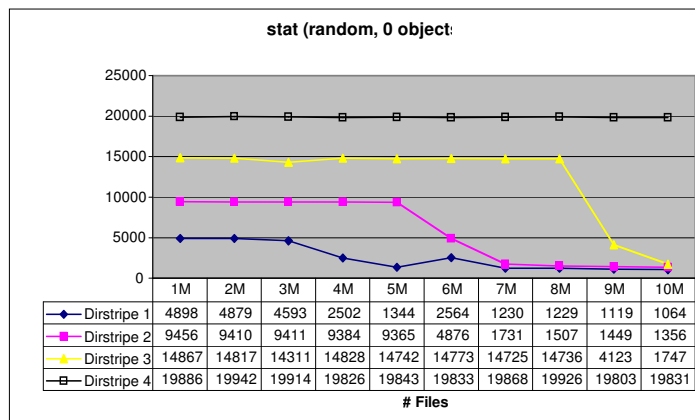
lustre™



CFS Cluster File Systems, Inc.™

Preliminary Results - stats

lustre™



CFS Cluster File Systems, Inc.™

Extremely large File Systems



- Remove limits (#files, sizes) & NO fsck ever
- Start moving towards “iron” ext3, like ZFS
 - Work from University of Wisconsin is a first step
 - Checksum much of the data
 - Replicate metadata
 - Detect and repair corruption
 - New component is to handle relational corruption
 - Caused by accidental re-ordering of writes
- Unlike “port ZFS approach” we expect
 - A sequence of small fixes
 - Each with their own benefits

Lustre Memory Client Writeback Cache



- Quantum Leap in Performance and Capability
- AFS is the wrong approach
 - Local server a.k.a. cache is slow
 - Disk writes, context switches
- File I/O writeback cache exists and is step 1
- Critical new element is asynchronous file creation
 - File identifier difficult to create without network traffic.
 - Everything gets flushed to the server asynchronously.
- Target: 1 client: 30GB/sec on mix of small / big files

Lustre Writeback Cache: Nuts & Bolts



- Local FS does everything in memory
- No context switches
- Sub tree locks
- Client capability to allocate file identifiers
 - Also required for proxy clusters
- 100% asynchronous server API
 - Except cache miss

Migration - preface to HSM...



OBJECT MIGRATION

- Intercept cache miss
- Migrate objects
- Transparent to clients



virtual OST

Hot Object Migrator



OSS2



OSS2

RESTRIPING POOL MIGRATION

- Coordinating client
- Control migration
- Control configuration change
- Migrate groups of objects
- Migrate groups of inodes

Virtual migrating MDS Pool

MDS Pool A

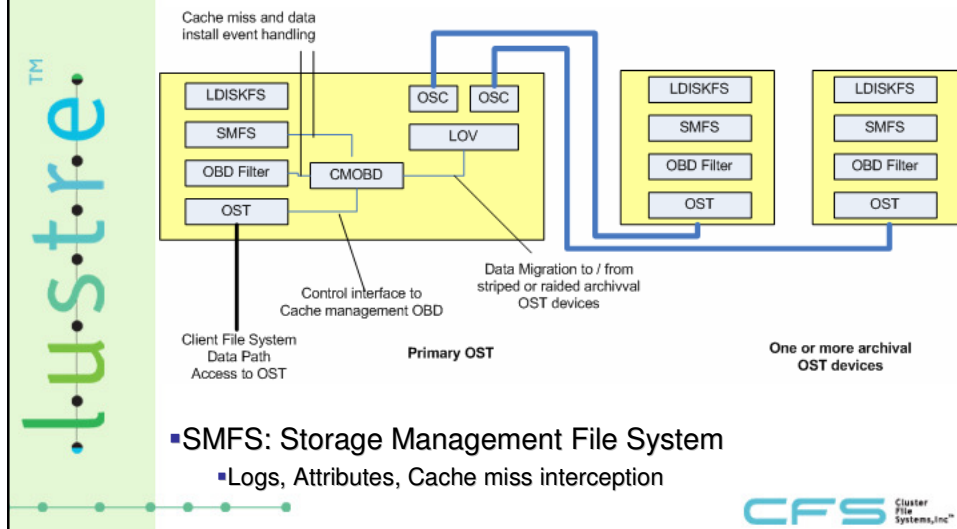
MDS Pool B

Virtual migrating OSS Pool

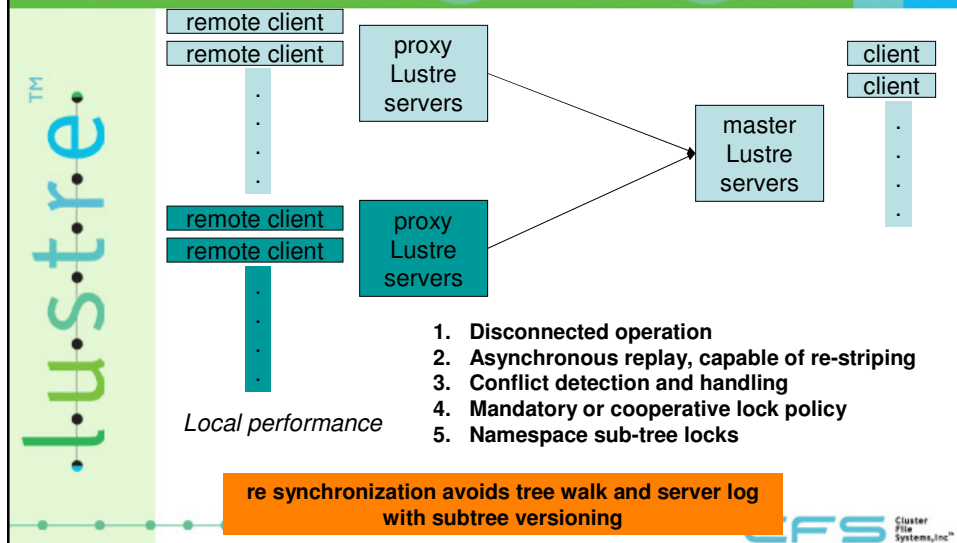
OSS Pool A

OSS Pool B

Cache miss handling & migration - HSM



Proxy Servers



Wide Area Applications



- *Real/Fast* grid storage
- Worldwide R&D environments in single namespace
- National Security
- Internet Content Hosting / Searching
- Work at home, offline, employee network
- Much much more...



Summary of Targets



	Total Throughput	Transactions per second	Client Throughput	System Scale
Today	~100 GB/s	9K ops/s	800 MB/s	300 Teraflops 250 OSSs POSIX
2008	1 TB/s	1M ops/s	30 GB/s	1 Petaflop 1000 OSSs Site wide, wb cache Secure
2010	10 TB/s			10 Petaflops 5000 OSSs Proxies, disconnected



Cluster File Systems, Inc.

lustre™

- **History:**

- Founded in 2001
- Shipping production code since 2003

- **Financial Status:**

- Privately held, self-funded, US-based corporation
- no external investment
- Profitable since the first day of business

- **Company Size:**

- 40+ full-time engineers
- Additional staff in management, finance, sales, and administration.

- **Installed Base:**

- ~200 commercially supported worldwide deployments
- Supporting the largest systems in 3 major continents [NA, Europe, Asia]
- 20% of the Top100 Supercomputers (according to the November '05 Top500 list)
- Growing list of private-sector users

CFS Cluster File Systems, Inc.™

Thank You!

CFS Cluster File Systems, Inc.™